Competition instructions

The competition on the **Kaggle** is launched and you should have received the invitations. If you have not received it, let me know.

Deadline for all the deliverables is March 27, at 23:59.

You will work in teams designed at the first seminar session and you will form these teams at the Kaggle website as well. Send me an email with the name of the team you will use on Kaggle at the first opportunity. Recall that the list of team members is posted in a Competition folder in Box.

At the end we opted for a single dataset that can be found at the UCI repository: http://archive.ics.uci.edu/ml/datasets/Covertype. It is not the most exciting dataset out there, but we had to choose relatively clear dataset so you do not have to work a lot on feature selection/transformation, since we did not discuss these machine learning topics in our classes.

You are free to use any classification method, regardless of whether it was used in class or not. Code will have to be implemented in R. There should be no code sharing between teams. To maximize the learning outcomes and to give you greater incentive to produce good code, after the competition is finished and grades delivered, the code used by each team to produce predictions (see points 3 and 4 below) will be made available to the whole class at Box. You are free to use any classification method, regardless of whether it was used in class or not. Code will have to be implemented in R. There should be no code sharing between teams. Winning code will be made available to the whole class.

Kaggle part of the competition will be used to determine the accuracy of your best classifier, however it is only one of the things you will have to deliver. This is because we want to see your reasoning behind choosing certain classifier over the others, how you optimized the parameters and how you were developing the code. The grading scheme is as follows:

1. **Accuracy** of the solution. This part will be completed on Kaggle In Class website. This will make **25%** of the competition grade. The best team will get full 25%, while other teams will get progressively smaller amounts.

2. **Report** showing the algorithms you have considered and tried out. We are interested in accuracy of your algorithm in absolute terms, but also relative, show us it is better than the alternatives you have tried out. Here you should also provide arguments for choosing one method over another, describe how you have chosen the parameters of your classifier and discuss advantages and limitations of your classifier. The report does not have to be extensive, but it has to have the elements mentioned above. We are open-minded about the format, it can be a PDF, a web page, R notebook or something else entirely. Do not think of it as a report only, but also as a presentation. Convince us why

is your solution the best one. This will make **25%** of the competition grade.

3. **Reproducibility** of your solution will make another **25%** of the grade. You will submit a code written in R that includes how exactly you have trained your final classifier, how you have used the training dataset posted on Kaggle and how it produced the predictions for the test set. Whatever manipulations you do with the initial training dataset to get to the one you finally used, it has to be done from R. We should be able to completely reproduce your solutions "with a single click". To achieve reproducibility you are free to use any technology you can think of, from a collection of R scripts and Readme files to R packages, makefiles to Docker containers, as long as you provide details on how to use it.

4. How **understandable** your code is for an external viewer. This will make another **10%** of the grade. When taking a look at the code you used for training your best classification algorithm and making predictions on the test data, we should be able to understand what you did. This usually depends on how commented the code is, how it is structured, how you named your variables etc.

5. Development of all the documents and the code related to the competition should be **under version control using Git**. This will make **15%** of the grade. It can be either a private or a public repository at some of the Git websites, such as Github, Bitbucket, Gitlab etc. At the end of the competition you will grant us access to the repository and we will check how actively you have used it to organize your work and produce the reports and the code for the competition. Bear in mind that we will reward proper usage of version control, not mindless committing after any tiny change to show your "activity". There are couple of ways to collaborate using Github, you will have to investigate it and choose one of them.

Work hard and have fun!
Hrvoje